ON THE COMPUTATION OF THE PROBABLE CORRECTNESS OF DIFFERENCES

By EDWIN G. BORING

One of two points of disagreement between Prof. Fernberger[1] and myself [2] is that, while I say that the significance of a difference between two limens or between a test-case and a norm must depend on the probable errors of each of the measures compared, he evidently believes that a significance may be reliably indicated when the probable errors are not available. And of course he is right, for practical instinct is not necessarily based on theoretical conviction. If—to take his instance of Grabfield[3]—we know that a faradic limen above 140 *B*-units has never been found in normal subjects, and that pathological subjects give limens from 120 to over 400 *B*-units (av. = about 200), then the clinician, who finds a threshold of 200, concludes with reasonable certainty that the case is pathological, although he may never know the P.E. of his subject, or of pathological subjects in general, or of the norm. How, then, does the clinician make his diagnosis?

Let us see first what theory requires. A norm is generally based upon many observations upon every one of many individuals. Thus it involves two modes of variability: an inter-individual mode, which is measured by the P.E. of the individual averages, and an intra-individual mode, which is measured by the average of the P.E.'s of the individuals. In theory we must take both into account, as we can by considering a new P.E. of all observations without regard to the individuals, *i.e.*, by taking the P.E. of a sort of 'group-individual.' In the diagnosis of a single subject, we must then consider both these modes of variability: we must find (1) the probability that we have not an instance that is exceptional for the particular individual, and (2) the probability that the individual's average is not merely an unusual case within the normal range of variability of individuals. Actually, we should have to find, by repeating the determinations of the threshold, the individual P.E. of our subject, and then throw that P.E. over against the P.E. which combines both modes of normal variability, before we could determine the probable correctness of the departure of our subject from the norm. If we can not take time to determine our subject's P.E., we may perhaps *assume* that his P.E. is no larger than the average P.E. for normal and pathological subjects, and then use some such measure in determining the P.E. of the difference. If the intra-individual P.E. is known to be fairly constant in all cases, we

[1] S. W. Fernberger, Concerning the Number of Observations Necessary for the Determination of a Limen, *Psychol. Bull.*, 14, 1917, 110.

[2] E. G. Boring, The Number of Observations upon which a Limen May be Based, *Amer. Jour. Psychol.*, 27, 1916, 315.

[3] G. P. Grabfield, Variations in the Sensory Threshold for Faradic Stimulation in Psychopathic Subjects, *Bost. Med. and Surg. Jour.*, 171, 1914, 883; a clinical article, unsatisfactory to the psychophysicist on account of the omission of data.

may perhaps make this assumption with validity; otherwise we must—there is no escape—determine the P.E. of the individual.

The instinctive diagnosis of the clinician means that he knows approximately what degrees of variability he may expect, and that instead of making nice calculations, he simply adopts a rule of refraining from positive diagnosis except when the difference is so great as to be for him unequivocally significant. To state that a faradic threshold greater than 175 *B*-units is "definitely pathological" is to imply a norm, an intra-individual and an inter-individual variability of that norm, and a pathological intra-individual variability. The clinician works implicitly where the psychophysicist works explicitly. Both may be right, but in the doubtful case the odds are with the psychophysicist.

On the other hand, unnecessary labor is always futile, and there are times when the "compromise between time and accuracy," which Fernberger urges, may gain for us more than it loses. If the psychophysicist can determine for the clinician a certain value as a *differentia* (like 175 *B*-units), which has, for known variabilities of the measures involved, a given probable correctness, then he saves the clinician's time. Moreover, there are ways in which he can save his own time.

In the first place, he may utilize that respectable mathematical notion of the negligible quantity. Fernberger laments that "the work of obtaining the probable error of this average threshold would be a very laborious affair." But perhaps it need not be determined; perhaps we may neglect it, that is to say, call it zero. The P.E. of a norm depends on the number of cases involved. If, for example, a norm is based on 50 times as many observations as is a given test-case, and if the variability of norm and test is the same, then, when the P.E. of the norm is taken as zero, the P.E. of the difference is altered by only one *per cent.* Since the norm usually does involve many more observations than the case to be compared with it, its P.E. can frequently be neglected. But one must neglect intelligently.

This principle of the negligible quantity can often be used in another manner. The norm, as we have seen, involves both intra-individual and inter-individual modes of variability. In comparing one individual with another, we work with the intra-individual variability; in comparing a group with another group (pathological cases with normal, perhaps), we take the inter-individual mode; in comparing an individual with a general norm, we use the combined measure of variabilty. If, however, either mode of variability is small with respect to the other, then, in those cases where the combined measure is needed, we may neglect the smaller. Presumably—we are not given the necessary data to make sure—the inter-individual variation in Grabfield's pathological cases is so large that it obscures a relatively small intra-individual variation, which we may therefore disregard.

In the second place, the psychophysicist, when his computation requires a P.E. which he does not know, may sometimes reason by analogy. Suppose that, from known averages and P.E.'s, he has found two or three groups, the averages of which are different, but not significantly different,—a frequent case, since significant differences in a common measure are rare. And suppose that the P.E.'s of these groups are practically the same. Then, if he is given the average of still another group without its P.E., he may assume that it has the same P.E. as the rest and compute the probable correctness of its difference from any one of the rest. If on this assumption he finds a significant difference, it is possible that he has assumed analogy on insufficient grounds; he should suspend judgment. If, on the other hand, he arrives with

his assumed P.E. again at a difference that is not significant, he will probably conclude that this is simply another case of an apparent difference due, not to the heterogeneity of the material, but to normal variability.

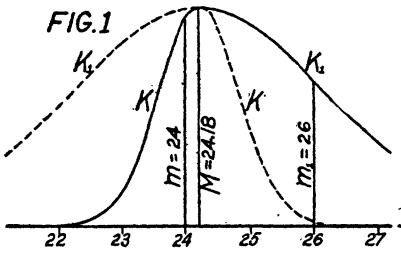
A second point of disagreement between Fernberger and myself concerns the generality of the application of the concept of numerical difference. Comparison, it seems to me, is as broad as science itself; in scientific experiment we change conditions and note the correlated change in a phenomenon, and this change is essentially a difference. Such a difference, when quantitatively expressed with available measures of precision, may conveniently be stated as a probability of difference (probable correctness). But Fernberger says: "We are also interested in the *amount* of that difference. It is true, for example, that we are interested in learning that the savage is more sensitive than the civilized man, but we are also interested in learning that the savage is *twice* as sensitive." And, undoubtedly, we might compute this probability, and also the probability that the savage is one and a half times as sensitive, and one and a quarter times, and so on indefinitely. In any such case we should always have two figures to keep in mind: the amount of the difference and the probability of that amount. If the total fact could be expressed by some single number, such as the probable correctness of the difference, surely an economy of thinking would be effected.

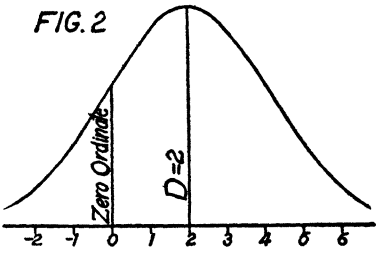
Let us be concrete. Suppose—to take a case of Henri's—that the average reaction-time in a set of 100 experiments is 24 hundredths of a second with a mean variation of 5, and that the average of another set of 25 experiments is 26 with a mean variation of 8. Is the difference between these two averages significant or due to chance? Our conclusion depends upon the six numerical values which we have just given, and we can not think in terms of six different values at once. We may, however, get rid of separate statements of number of cases by writing the P.E.'s of the means, so that we now have but four numbers to keep in mind; $24 \pm .423$ and 26 ± 1.352. We often wish to relate two such pairs of values without knowing how to do it; hence we had better simplify again by computing the difference, which is 2, and its probable error, which is 1.417. Now with only two terms we are getting within thinkable compass; and Fernberger has in them an expression for the amount of the difference and its probability. But we may prefer to simplify still further. Suppose someone should ask us whether this difference of 2 ± 1.417 indicates greater or less homogeneity than a difference, say, of 5 ± 2.8; we probably could not say. As a final simplification we may, therefore, take the ratio of the difference to its P.E.—in this present case, 1.411. Now we have reduced a six-fold statement to a single value. We have not really lost the amount of the difference, because all that we ever had was a difference with a certain probable error upon it, and the probable correctness reflects both the amount of the difference and its P.E. After we have convinced ourselves that a given disparity is significant, we may wish again to take account of its amount; but in so doing we imply a prior interest in its probable correctness. We are, as I have said, "interested in differences *between* limens;" first in the probable correctness and then, sometimes, in the amount.[4]

[4] In general I believe that Fernberger and I are in accord. I did not oppose his recommendations. I sought merely to point out some further implications of his data and to suggest the method of dealing with them.

What I have called the probable correctness of a difference is the probability that a given difference will not be negative, that is to say, that there will be disparity in the same direction as in the observed measures.[5] I have advocated this as the simplest possible measure of disparity and as one that is easily computed.

The question has been raised as to the relation of this "probable correctness" to the "probability that a difference is not due to chance," for which Henri has published a formula.[6] Since the nature of Henri's formula is not easily understood, it may be well to take this opportunity to explain it.





Let us take the case of the reaction times (Henri, p. 159) mentioned above and let us construct a figure in explanation (Fig. 1). We have two means, $m = 24$ and $m_1 = 26$, which we may lay off on our abscissa. We assume tentatively that the two means represent homogeneous data. We now find a weighted average, M, between the two means, which depends on the measure of precision of each mean and the number of cases involved in each; Henri gives the formula (p. 155). M=24.18. If the two means are really homogeneous, then this weighted average is the most probable value of the measure in question and should represent the median of a normal curve of distribution. We can draw such a curve about M if we know its measure of precision. If we regard m as a chance deviate from M, then we may determine the measure of precision with respect to which m, as a mean, varies (p. 156, top); and with this measure of precision we may draw the curve K, of which m becomes an ordinate. Now by means of formula

[5] It can be found, when we know the ratio of the difference to its P.E., from a table of the probability integral; *op. cit.*, 317. In the Cornell Laboratory we have a negative with the requisite equations and a graph of the probability integral, which make the finding of a probable correctness a simple matter. We can furnish blue prints from this negative at a very slight cost.

[6] V. Henri, Quelques applications du calcul des probabilités à la psychologie, *L'année psychol.*, 5, 1898, 153. I am indebted to Dr. L. T. Troland for pointing out the asymmetry of Henri's formula. It is in an attempt to meet certain difficulties that he raised that I add this discussion.

Henri's paper is both obscure and confused by inaccuracies. The following errata should be noted: (1) P. 156, formula (2): for "2_1", read "n_1". (2) P. 159, line 5: for "v", read "v_1". (3) P. 159, line 21: for "3.7", read "3.57". (4) P. 159, last two lines: for "$\frac{1}{89} = 0{,}01$", read "$\frac{100}{281} = 0{,}35$"; and for "0,02", read "0,22". The last error is a miscalculation which almost reverses a conclusion.

(2), (3), or (4) and a table of the probability integral, we find the area included between M and m. Thus, for $m = 24$, $t = .1988$ and the table shows that this area is 22% of the total area to the left of M. (Henri miscalculates this area as 2%!) This area is a measure of the amount that the mean, m, deviates from the theoretical mean, M, which, on the assumption that m and m_1 are homogeneous, is the most probable true value of the measure. But the farther m lies from M the less likely is it that m and m_1 (m_1 by being included in M is the cause of the disparity between m and M) are homogeneous. Thus Henri uses the area between m and M—22% in this case—as a measure of the probability that the difference is due to a determinate cause and not to chance. Conversely the probability that the difference is due to chance is 78%. Here Henri leaves us.

The difference that Henri has been measuring is not, as he implies, the difference between m and m_1; it is the difference between m and M. We can apply the same procedure to m_1 that we did to m. In this case we erect on M the curve K_1, of which the measure of precision is found from the data for m_1. K and K_1 will be different so long as the number of cases in m and m_1 or the measures of precision of m and m_1 are different. The probability that the difference between m_1 and M is due to a determinate cause is 64%. (Henri does not calculate this case.)

Thus it appears that the probability of an operative cause is 22% in the case of m and 64% in the case of m_1, when each is compared with the theoretically most probable mean, M. Can we get a single measure, as Henri does not, for the difference between m and m_1? If m and m_1 were really homogeneous, they would lie on the same curve and we could take the area between them. But, although Henri has considered the two as homogeneous, he has used different measures of precision for each of them, on the assumption (an error, surely) that the proper measure should depend solely on the one variant under consideration, and not upon both of the supposedly homogeneous variants. This mistake is theoretically fatal.[7] Nevertheless, if we want merely a practical measure, we may take the asymmetrical curve made up of K on the left and K_1 on the right, and (considering the areas on either side of M as equal) find the *per cent.* of the total area included between m and m_1. It is 43%. Such a value is not mathematically defensible, but it gives in practice results consistent with the values of "probable correctness," although smaller. It is the only way—so far as I can see—in which Henri's formula can be used in all cases to indicate the significance of a difference.

Fig. 2 shows the "probable correctness" of the difference in this same case of the reaction-times. The actual difference, D, is assumed to vary along a normal distribution curve, for which D is the most probable value, and for which the P.E. of D is the measure of variability. The probable correctness of the difference is the probability that D, in varying, will not assume a negative value, but will always represent a disparity in the same direction. This probable correctness

[7] On the other hand, some such incompatibility is essential to this mode of reasoning. When Henri assumed that m and m_1 were homogeneous, he assumed implicitly that one is just as often greater as less than the other,—the very sort of conclusion that his method aims to give. His formulation of the problem is fundamentally impossible; to think of two means as variants on a single curve is to deny that they have upon the curve fixed places, which would make it possible to determine a distance or an area between them.

is, therefore, the part of the total area under the curve which lies to the right of the zero-ordinate: in this case, 83%.

Since 50% represents pure chance, we may say that a probable correctness of 83% is .66 from pure chance to certain cause. The corresponding value that we got by extending Henri's method was .43. The reason for the discrepancy (.66 and .43) appears in a consideration of the two Figg. In Fig. 1 the difference m_1—m would be reversed if m, fell to the left of m or if m fell to the right of m_1. The chance of m_1 falling to the left of m is 39%; the chance of m falling to the right of m_1 is 18%. The chance of the difference being reversed in the one or the other of these ways is thus 57%, that is to say, the area not included between m and m_1. One is tempted to say, then, that the area between m and m_1 (43%) must give the probable correctness of the difference. This, however, is not the case, for m and m_1 ought to be taken as varying simultaneously. There would then be some cases of a reversed difference when both lay between their present values. Since the extension of Henri's formula omits these cases, the probable correctness (probability that the difference is not negative) must be a value greater than 43%. We have found it by the other method to be 66%, provided we take pure chance as zero, as we do in Henri's computations.[8]

It appears, then, that we may extend Henri's method to compute the probability of difference. This extended method and my own method indicate approximately the same thing; but Henri's method involves an unwarrantable mathematical assumption, it fails further to consider the two means as simultaneously variable (thereby giving values too low), and it requires approximately twice as many operations for the determination of a 'probable correctness.' What we want, I take it, is a formula that is both adequate and simple.

[8] We can never, of course, come out with anything more than a probability. To ask whether or not a difference is significant is to talk common sense and mathematical nonsense. Mathematically we can indicate any degree of probable correctness between the limits of complete certainty and pure chance; and differences with various degrees of probable correctness are variously significant. A line between what is significant and what is not can be drawn only by some arbitrary convention. We may place it according to our own personal convictions or with respect to a consensus of scientific usage; or we may take into account considerations of symmetry, as Henri does, and call significant everything that is nearer certainty than pure chance. In this last case probable correctnesses over 75% would be designated as significant. Psychologists would probably put the division at 85 to 95%. The question is one for scientists and not for science, and it can never be permanently settled.